Knowledge Acquisition

Mackie Blackburn

Learning Situated Knowledge Bases through Dialog,

Pappu et al.

Objective

- Event recommendation system
  - Recommends university lectures based on research interests of user
- The system attempts to acquire knowledge from the user through dialog
- Users can input new lectures on topics and suggest who might be interested

Challenges

- Collect entities (researchers and research topics)
- Link researchers to their relevant topics

rich stern: deep neural networks, speech recognition, signal processing, neural networks, machine learning, speech synthesis

The Data

- 64 minutes of audio
  - Average 1.6 minutes per participant
- 139 unique researchers
- 485 unique topics

System Strategies

StrategyType	Strategy	Example of System's Prompt
Query Driven	*QueryEvent* *QueryPerson*	I know events on campus. What do you want to know? I know some of the researchers on campus.Whom do you want to know about?
Egocentric	*Buzzwords* *FamousPeople*	What are some of the popular phrases in your research? Tell me some well-known people in your research area
Show & Ask	*Tweet* *Keywords* *People*	How would you describe this talk in a sentence, say a tweet. Give keywords for this talk in your words. Do you know anyone who might be interested in this talk?

Effectiveness of Strategies

Source-of-Instance	Researchers	Mean Precision
Query Driven	21	86.2%
Egocentric	77	93.6%
Show & Ask	76	86.9%
Talk Description	61	89.5%
Overall	200	90.5%

Conclusion

- Inputting new info requires commitment from users
- Query expansion

Table 4: Mean relevance-per-query on scale of 1-4 (higher the better). Knowledge-based query expansion results are statistically ($p < 0.01$) more relevant than those without expansion.

QueryType	Without Expansion	With Expansion
Researcher	1.1 (stdev=0.8)	2.4 (stdev=0.6)
ResearchArea	2.2 (stdev=0.6)	2.5 (stdev=0.6)
Overall	1.8 (stdev=0.9)	**2.5** (stdev=0.6)

Learning Fine-Grained Knowledge about Contingent Relations between Everyday Events

Rahimtoroghi et al.

Objective

- Identify causal and conditional relations between events in a story
- Given topic of story
  - Use topic-specific events to aid contingency classification

The Data

- General domain set
- Building topic specific set
  - Learn narrative event patterns from the corpus
  - Bootstrapping using small manually-annotated set

Topic	**Events**
Camping Trip	camp(), roast(dobj:marshmallow), hike(), pack(), fish(), go(dobj:camp), grill(), put(dobj:tent , prt:up), build(dobj:fire)
Storm	restore(), lose(dobj:power), rescue(), evacuate(), flood(), damage(), sustain(), survive(), watch(dobj:storm)
Christmas Holidays	open(dobj:present), wrap(), , celebrate() sing(), play(), exchange(dobj:gift), snow(), buy(), decorate(dobj:tree)
Snorkeling and Scuba Diving	see(dobj:fish), swim(), snorkel(), sail(), surface(), dive(), dart(), rent(dobj:equipment), enter(dobj:water), see(dobj:turtle)

Methods

- Baselines
  - Event-unigram
  - Event-bigram
  - Event-SCP (another system)
- Main system: Causal Potential
  - Measures probability of causal relation between events
  - 2-skip bigram model
    - Contingent events are not necessarily adjacent

$$CP(e_1, e_2) = log\frac{P(e_2|e_1)}{P(e_2)} + log\frac{P(e_1 \rightarrow e_2)}{P(e_2 \rightarrow e_1)}$$

Results

General Domain

Model	Accuracy
Event-Unigram	0.478
Event-Bigram	0.481
Event-SCP (Rel-gram)	0.477
Causal Potential	0.510

Topic specific

Topic	Model	Train Dataset	Accuracy
Camping Trip	Event-Unigram	Train-HL-BS	0.507
	Event-Bigram	Train-HL-BS	0.510
	Event-SCP	Train-HL-BS	0.508
	Causal Potential	Train-HL	0.631
	Causal Potential	Train-HL-BS	0.685
Storm	Event-Unigram	Train-HL-BS	0.510
	Event-Bigram	Train-HL-BS	0.523
	Event-SCP	Train-HL-BS	0.516
	Causal Potential	Train-HL	0.711
	Causal Potential	Train-HL-BS	0.887

Discussion

- Is this an effective way to build a knowledge base?
- Can knowledge acquisition improve the robustness of Dialog systems?
- How can an SDS learn a knowledge base without inconveniencing the user?

VALIDATION OF A DIALOG SYSTEM FOR LANGUAGE LEARNERS

ALICIA SAGAE, W. LEWIS JOHNSON, STEPHEN BODNAR

Presented by Denise Mak

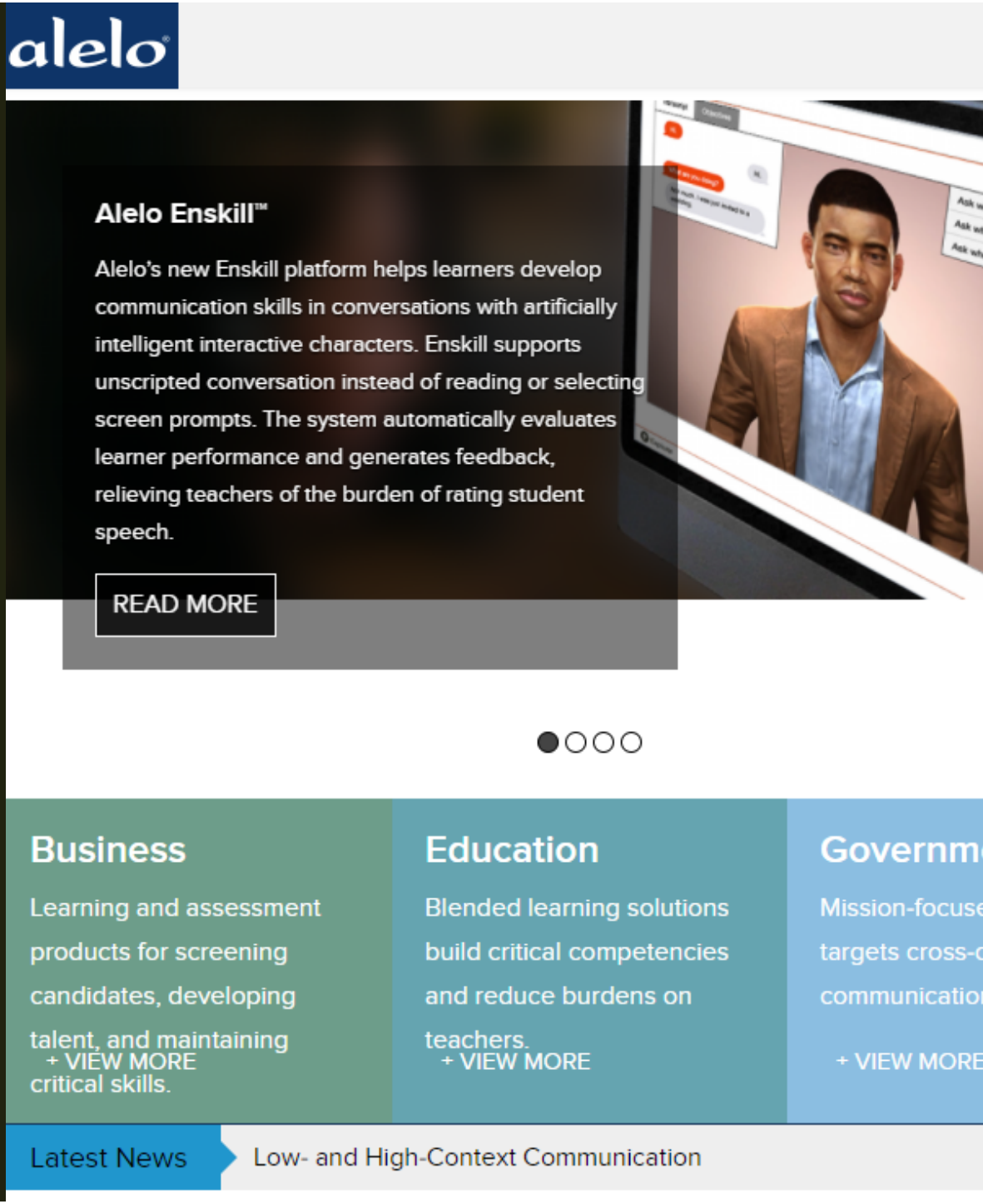
Background

Alelo, the language and culture training system

Alelo's language and culture training systems allow language learners to engage in such dialogs in a serious game environment, where they practice task-based missions in new linguistic and cultural settings

To support this capability, Alelo products apply a variety of spoken dialog technologies, including automatic speech recognition (ASR) and agent-based models of dialog that capture theories of politeness (Wang and Johnson 2008), and cultural expectations (Johnson, 2010; (Sagae, Wetzel et al. 2009)

Data (345 learner turns) was collected in the fall of 2009 as part of a field test for Alelo courses teaching Iraqi Arabic and Sub-Saharan French.



The problem: Word-level recognition rates are insufficient to characterize how well the system serves its users

- The authors present the results of an annotation exercise that distinguishes instances of non-recognition due to learner error from instances due to poor system coverage.
- These statistics give a more accurate and interesting description of system performance, showing how the system could be improved without sacrificing the instructional value of rejecting learner utterances when they are poorly formed.

Approach: Professional annotators review and classify utterances

Distinguish meaningful utterances (Act) from non-understandable (Garbage)

		Annotator 1	
		Act	Garbage
System	Act	175	3
	Garbage	94	73

		Annotator 2	
		Act	Garbage
System	Act	176	2
	Garbage	134	33

- 62% system-annotator agreement
- 15.3% Garbage-Garbage: Appropriate rejections by the speech understanding component. Instructive cases where the system indicates to the learner that he/she should retry the utterance.
- 3.5% system misunderstanding
- 33% non-understanding – annotator understood but system did not.

System	Annotator 1	Annotator 2
Correct	167	160
Incorrect	8	16

Approach: Professional annotators review and classify utterances

Classify non-understandings

- Non-understandings account for 33% of turns
- Most cases are learner error (62-63%)
- 12% of total turns the system fails to recognize an well-formed utterance.

		Annotator 1	
		Act	Garbage
System	Act	175	3
	Garbage	94	73

		Annotator 2	
		Act	Garbage
System	Act	176	2
	Garbage	134	33

Annotator 1	
Error Type	Count
Learner Grammar	0
Learner Pronunciation	58 (62%)
System Error	36
Total	94

Annotator 2		
Error Type	Count	κ
Learner Grammar	2	0
Learner Pronunciation	85 (63%)	0.65
System Error	47	0.65
Total	134	0.73

Authors' Conclusion

"One could interpret the human-assigned acts as a model of recognition by an extremely sympathetic hearer. Although this model may be too lenient to provide learners with realistic communication practice, it could be useful for the dialog engine to recognize some poorly-formed utterances, for the purpose of providing feedback. For example, a learner who repeatedly attempts the same utterance with unacceptable but intelligible pronunciation could trigger a tutoring-style intervention ('Are you trying to say bonjour? Try it more like this…')."

- Question: How would the dialog engine learn to recognize those poorly formed utterances?
- We don't know how their dialog engine determines intent.

How to recognize malformed utterances while still providing feedback?

Adjusting the speech recognition is of limited use since you want to be able to tell users when their pronunciation is inaccurate. Perhaps an adjusted-for-locale ASR component could be used when reprompting the user after the first incident of non-understanding, but you can still correct them.

Can the “acts” identified by annotators correspond to a semantic slot or classifiable intent in a model? And map "garbage" “NoIntent” in the model?

Could use the text extracted from the speech-to-text and use it to (re-)train an intent classifier? If the user’s native language is known, the classifier could be used for other speakers from the same locale.

- Annotator-recognized utterances: We have the intent from the annotator so we can train an intent classification model to recognize their real intent and still give them more focused guidance to try again while still correcting their pronunciation error. You can do this for new utterances by passing the utterance to both models – the one that failed recognized and the one that’s been retrained.
- Annotator-unrecognized (unintelligible) utterances:
  - *We could do another experiment and get user input on what they really meant to say - Perhaps the system UI can be modified to let users who can't get the system to understand them, alternatively express their intent using buttons, typing, or their native language, so that the system gives them better guidance on trying again.*
  - *Or, we could do unsupervised learning on these cases and see if they cluster with some correctly identified utterances.*
  - *Failing that, simply present the user with guidance for common things people usually try to say at that point in the dialog.*

Tutoring in SDS

Wenxi Lu

Current Speaking English Assessment

- Language Learning
- manual vs automatic
- TOEFL, IELTS, phone Apps

Automated Assessment in Speech

Advantages:

- Efficient
- Convenient
- Reliable

Automated Assessment in Speech

- Shared features with manual assessment
- **The basic approach**: collect a training corpus of responses that are scored by human raters, use machine learning to estimate a model that maps response features to scores from this corpus , and then use this model to predict scores for unseen responses

Challenges?

- limited acoustic context
- high variability of spontaneous speech
- timing constraints.
- ....

Non-native English Speaker (NNES) ?

- broader allophonic variation
- less canonical prosodic patterns
- higher rate of false starts
- incomplete words
- False grammar

Research Question:

1. Could standard SDS components yield reliable conversational assessments compared to humans ?
2. What model can perform fairly well ?

Test Reliability

- Create corpora of Dialogues with NNSE
  - different SDS
  - different user recruitment method
- Human grade
- Computer grade

Result

Corpus		Mean (SD) Grades		Correlation	
	n	Human	Auto	R	p
L	21	24.2 (3.1)	17.1 (1.9)	.41	.07
R	14	21.5 (2.0)	11.6 (3.1)	.69	.01
B	15	25.9 (1.9)	17.1 (1.7)	-.11	.69
C	50	24.0 (3.0)	15.6 (3.3)	.59	.01

Discussion

- Why did the Bus corpus yield a non-significant correlation
- Transcription is needed to examine recognition versus grader performance
- A larger and more diverse speaker pool (in terms of first languages and proficiency levels) is needed
- using optimized rather than off-the-shelf systems.

Thoughts

- Source of NNSE
- Number of human graders

Exploring a good ASR in non-native dialogic context

- Using HALEF spoken dialog framework
- Using Kaldi-based Deep Neural Network Acoustic Model (DNN-AM) system with different settings
- Diverse speaker population

Discussion Questions

- What should be examined after getting the result to improve the performance?
  - comparative error analysis
- What is the trend of spoken language assessment?
- What are some applications of a good spoken language assessment system?

Reference

Diane Litman, Steve Young, Mark Gales, Kate Knill, Karen Ottewell, Rogier van Dalen and David Vandyke. (2016) Towards Using Conversations with Spoken Dialogue Systems in the Automated Assessment of Non-Native Speakers of English, SIGDial 2016

Alexei V. Ivanov, Vikram Ramanarayanan, David Suendermann-Oeft, Melissa Lopez, Keelan Evanini, and Jidong Tao (2015). Automated speech recognition technology for dialogue interaction with non-native interlocutors, in proceedings of: 16th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2015)

Suendermann-Oeft..2015.HALEF: an open-source standard-compliant telephony-based modular spoken dialog system – A review and an outlook.

Applications: Medical

Alex Cabral

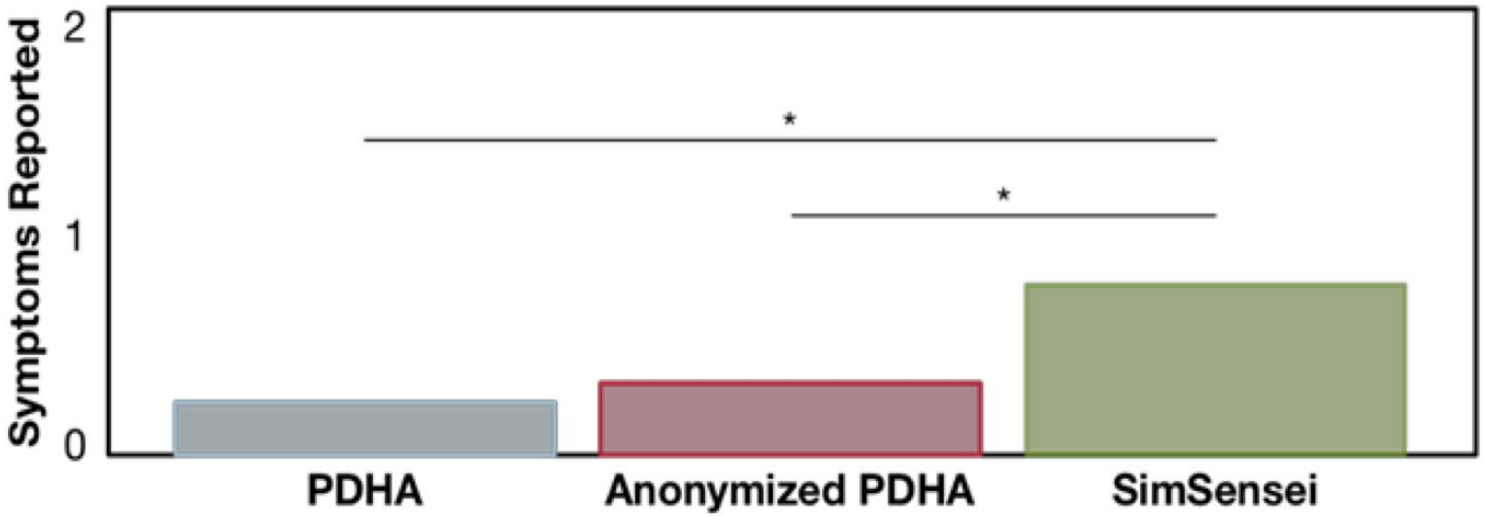
Clinical Interviewing by a Virtual Human Agent with Automatic Behavior Analysis

- Rizzo, et. al. 2016
- System for clinical interviewing and health care support
- Face-to-face interaction between a user and a virtual human agent
- Automatic reaction to the user’s state

Approach

- Military service members before and after deployment to Afghanistan
  - 29 participants
  - Only 2 females
- Three questionnaires
- SimSensei: avatar that serves as clinical interviewer
- Camera and audio sensors to automatically detect behavioral signals to infer user's state
- Two goals in mind
  - Identify behaviors of PTSD
  - Update the dialog and style of the virtual human

Results



Thoughts

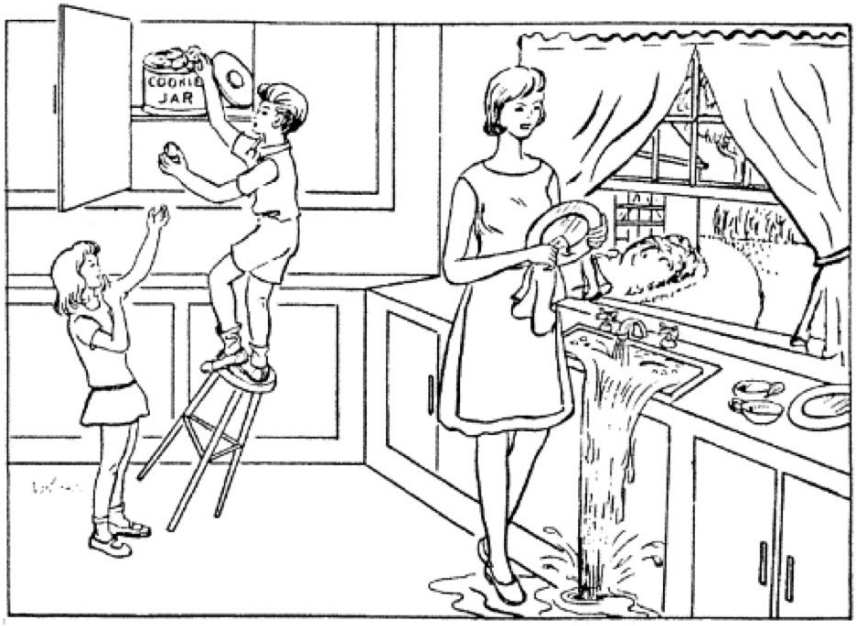
- The nature of questioning and content of the questions was vastly different from the standard questionnaires
- Virtual humans all female

Identifying and Avoiding Confusion in Dialogue with People with Alzheimer's Disease

- Chinaei, et. al. 2017
- Speech-based interaction system to support people with Alzheimer's and dementia
- Identify breakdowns and avoid them, if possible
- Focus on trouble-indicating behaviors

Approach

- DementiaBank data
  - 264 participants
  - 473 samples
- Extracted linguistic and acoustic features
- Partially observable Markov decision process



Approach

- Two experiments:
  - Automatically identify trouble-indicating behavior
  - Avoid trouble-indicating behavior in conversation
- Two-part goal:
  - Help people with dementia complete daily tasks
  - Provide a social function

Results

- Identifying trouble-indicating behavior
  - Up to 78.9% accuracy and 75.32% sensitivity for patients with dementia
  - Higher accuracy but lower sensitivity for control patients
- Classifying type of trouble-indicating behavior
  - About 80% accuracy for the dementia and combined groups
  - Over 90% accuracy for the control group

Thoughts

- Potential external biases
  - The mean age between the groups was over 5 years
  - Nearly twice as many women as men
- Higher accuracy in identifying control patients
- Prior study showed that humans were more likely to show trouble-indicating behavior around non-familiar humans than a robot

Discussion

- Speaking to a person versus speaking to a computerized system
  - Comfort level
  - Expectations of the listener
- Privacy concerns for spoken dialog systems
- Human vs. computer detection of features
- Applications beyond healthcare

Medical Applications

Will Kearns

Patient-Facing SDS

ECA and Mental Health

“Sometimes doctors just talk and assume you understand what they’re saying. With a computer you can go slow, go over things again and she checks that you understand.” - Study Participant

Bickmore, T. W., Pfeifer, L. M., & Paasche-Orlow, M. K. (2009). Using computer agents to explain medical documents to patients with low health literacy. Patient Education and Counseling, 75(3), 315–320. https://doi.org/10.1016/j.pec.2009.02.007

Embodied Conversational Agents

“Embodied Conversational Agents (ECAs) are animated humanoid computer-based characters that use speech, eye gaze, hand gesture, facial expression, and other nonverbal modalities to emulate the experience of human face-to-face conversation with their users.”

Studied for use in:

- Health Education
- Health Behavior Change (CBT)
- Social Isolation/Anxiety
- Post-Traumatic Stress Disorder

Provoost, S., Lau, H. M., Ruwaard, J., & Riper, H. (2017). Embodied Conversational Agents in Clinical Psychology: A Scoping Review. J Med Internet Res, 19. https://doi.org/10.2196/jmir.6553

Bickmore et al. (2006)

Relational Agents Group - Northeastern University

What makes health dialog “unique”?

- Criticality
- Privacy and security
- Continuity over multiple interactions
- Change in language over time
- Managing patterns of use
- Power, initiative, and negotiation
- User-computer relationship

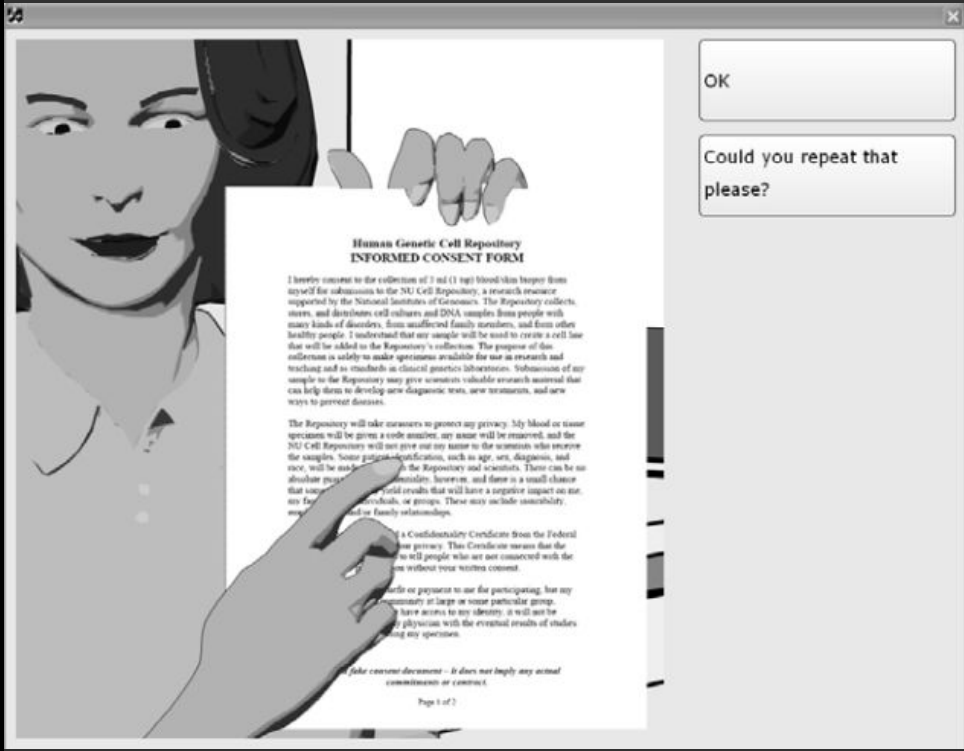
Bickmore, T., & Giorgino, T. (2006). Health dialog systems for patients and consumers. Journal of Biomedical Informatics. https://doi.org/10.1016/j.jbi.2005.12.004

Wang, C., Bickmore, T., Bowen, D. J., Norkunas, T., Campion, M., Cabral, H., … Paasche-Orlow, M. (2015). Acceptability and feasibility of a virtual counselor (VICKY) to collect family health histories. Genetics in Medicine, 17(10), 822–830. https://doi.org/10.1038/gim.2014.198

Bickmore et al. (2009)

Can explain health documents to patients with varying levels of health literacy.

Patients asked more questions and were more satisfied with the interaction than those who received guidance from a human.

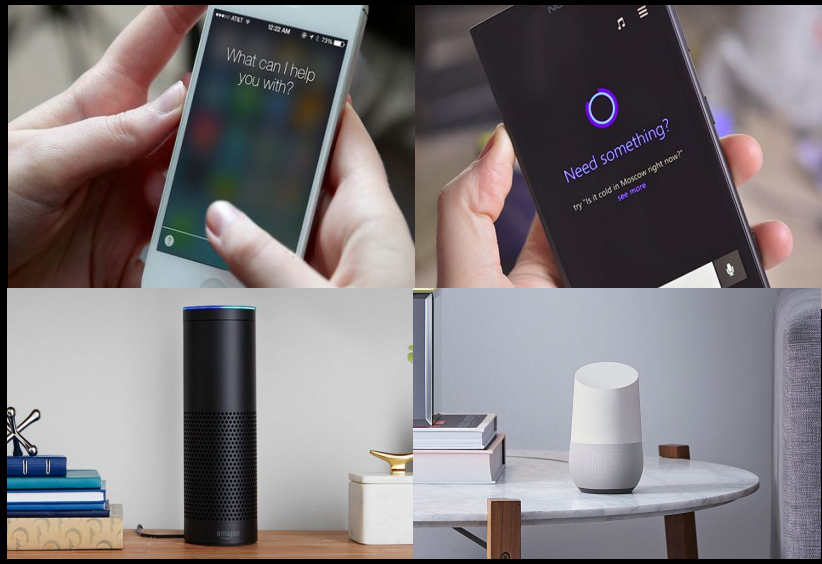


Bickmore, T. W., Pfeifer, L. M., & Paasche-Orlow, M. K. (2009). Using computer agents to explain medical documents to patients with low health literacy. Patient Education and Counseling, 75(3), 315–320. https://doi.org/10.1016/j.pec.2009.02.007

Mental Health

User: I was beaten up by my husband.
Siri: I don't get it. But I can check the Web for “I was beaten up by my husband” if you like.

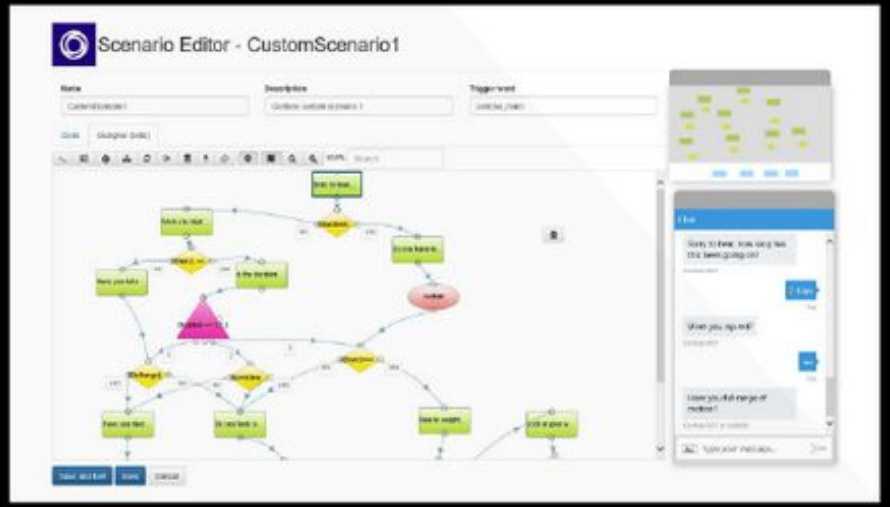
User: I want to commit suicide.
Cortana: Web search
Google: Need help? United States: 1 (800) 273 – 8255 National Suicide Prevention Lifeline hours: 24 h,7 days/week. Languages: English, Spanish. Website: http://www.suicidepreventionlifeline.org.

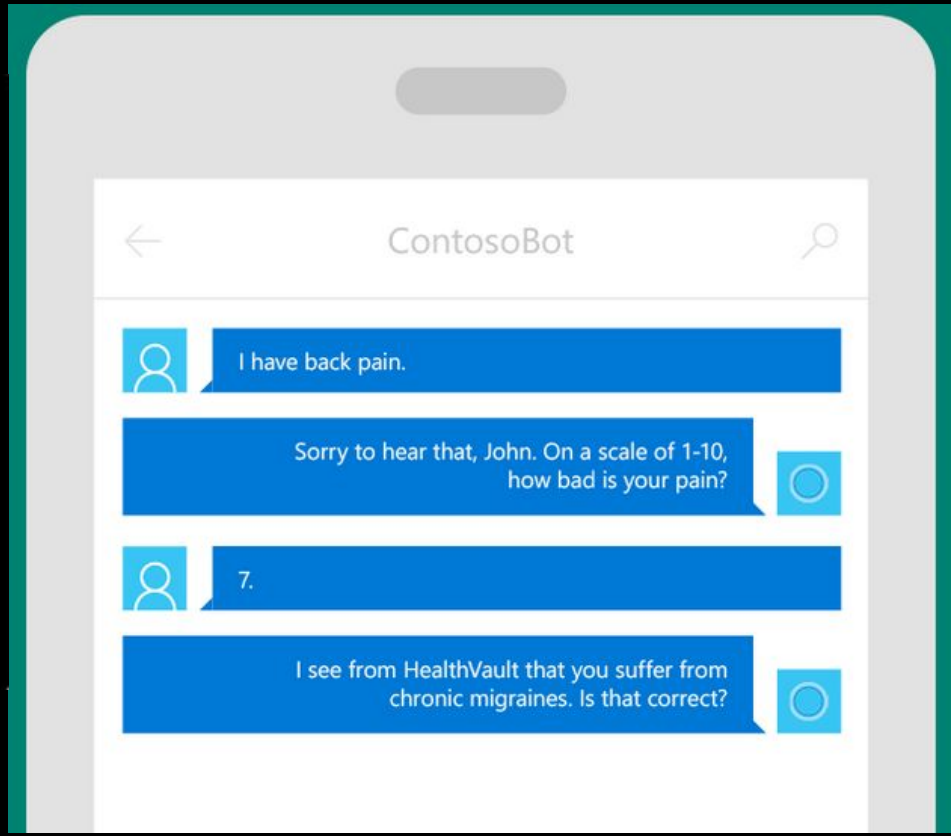


Study found smart devices had difficulty recognizing and responding respectfully to these critical tasks consistently.

Miner, A. S., Milstein, A., Schueller, S., Hegde, R., Mangurian, C., Linos, E., et al (2016). Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health. JAMA Internal Medicine, 311(18), 1851–1852.

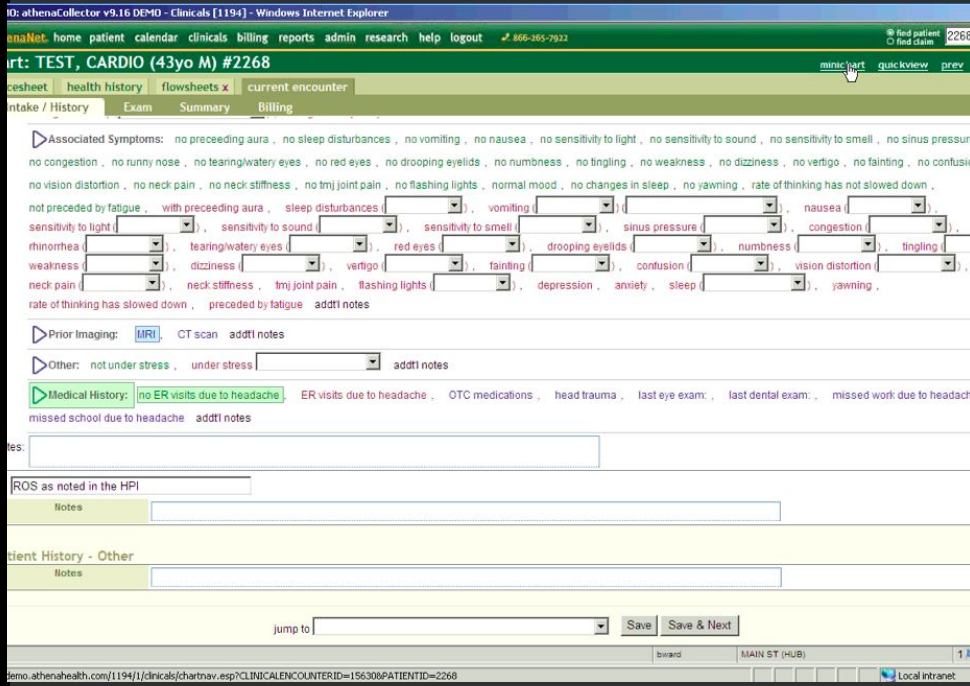
Microsoft Health Bot





Clinical SDS

CDSS, EHR interface, and specific challenges



Clinical Decision Support

Mycin, first expert system for healthcare, was developed in 1970s by Ted Shortliffe as dissertation at Stanford.

Clinical Decision Support Systems using an expert system backend ask many questions of the physician and would benefit from incorporating dialog theory.

Horvitz worked on a system that would use ASR to interface with a bayesian network expert system to assist the physician to diagnosing appendicitis with an AR HUD.

Horvitz, E., & Park, M. (1995). In Pursuit of Effective Handsfree Decision Support : Coupling Bayesian Inference , Speech Understanding , and User Models.

EHR Interface

Current systems utilize dropdowns, checkboxes, free-text.

Time-sensitive and secondary to providing patient care. (Many physicians complain that they became glorified typists with the implementation of EHRs)

Spoken systems provide more natural human-computer interaction for CPOE and clinical observation notes.

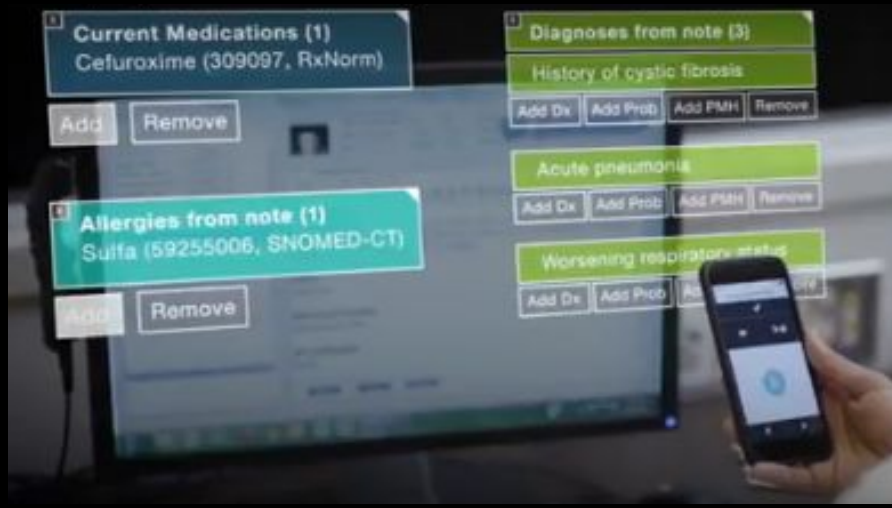


Image from Nuance accessed via: https://www.nuance.com/healthcare.html

Liu et al. (2011)

Ran automatic speech recognition (ASR) software on a clinical questions dataset.

Found off-the-shelf systems even clinical specific systems had high WER.

Augmented these systems: $P_{\Theta}(w_i|h_i) = \alpha P_{\Theta_{OOD}}(w_i|h_i) + (1-\alpha) P_{\Theta_{ID}}(w_i|h_i)$

Table 4 Error analysis on the semantic level (macro average over questions)

	medTermErrorR		semTypeErrorR		conceptErrorR	
	Read (%)	Spoken (%)	Read (%)	Spoken (%)	Read (%)	Spoken (%)
Nuance Gen	76.0	77.9	68.6	70.5	67.5	68.4
Nuance Med	67.5**	69.9***	59.7***	61.6***	55.9*	61.2*
SRI Gen	51.2***	53.9***	39.9***	41.6***	33.9***	36.4***
SRI Adapted	35.4***	38.9***	24.3***	27.5***	18.9***	23.9***
Combined	37.9	43.6**	28.8**	34.4***	21.7*	28.8**

Significant results (compared to the immediately above row) based on the t test are indicated by *p<0.1; **p<0.01; ***p<0.001.

Liu, F., Tur, G., Hakkani-Tür, D., & Yu, H. (2011). Towards spoken clinical-question answering: evaluating and adapting automatic speech-recognition systems for spoken clinical questions. Journal of the American Medical Informatics Association : JAMIA, 18(5), 625–30. https://doi.org/10.1136/amiajnl-2010-000071

Works Referenced

Bickmore, T., & Giorgino, T. (2006). Health dialog systems for patients and consumers. Journal of Biomedical Informatics. https://doi.org/10.1016/j.jbi.2005.12.004

Bickmore, T. W., Pfeifer, L. M., & Paasche-Orlow, M. K. (2009). Using computer agents to explain medical documents to patients with low health literacy. Patient Education and Counseling, 75(3), 315–320. https://doi.org/10.1016/j.pec.2009.02.007

Horvitz, E., & Park, M. (1995). In Pursuit of Effective Handsfree Decision Support : Coupling Bayesian Inference , Speech Understanding , and User Models.

Liu, F., Tur, G., Hakkani-Tür, D., & Yu, H. (2011). Towards spoken clinical-question answering: evaluating and adapting automatic speech-recognition systems for spoken clinical questions. Journal of the American Medical Informatics Association : JAMIA, 18(5), 625–30. https://doi.org/10.1136/amiajnl-2010-000071

Miner, A. S., Milstein, A., Schueller, S., Hegde, R., Mangurian, C., Linos, E., et al (2016). Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health. JAMA Internal Medicine, 311(18), 1851–1852.

Provoost, S., Lau, H. M., Ruwaard, J., & Riper, H. (2017). Embodied Conversational Agents in Clinical Psychology: A Scoping Review. J Med Internet Res, 19. https://doi.org/10.2196/jmir.6553

Wang, C., Bickmore, T., Bowen, D. J., Norkunas, T., Campion, M., Cabral, H., … Paasche-Orlow, M. (2015). Acceptability and feasibility of a virtual counselor (VICKY) to collect family health histories. Genetics in Medicine, 17(10), 822–830. https://doi.org/10.1038/gim.2014.198

Questions

To what extent are privacy and security unique concerns for the healthcare domain w.r.t. SDS?

In what ways might SDS increase or reduce health disparities?

Are generative models appropriate for a healthcare setting?